PathMaster: A Web-Accessible Cell Image Database Indexed by Mathematical Descriptors and Supported by Parallel Computation

Mark A. Shifman, MD, PhD, Mark Mattie, MD, PhD, Frederick G. Sayward, PhD, Perry L. Miller, MD, PhD, Yale University School of Medicine, New Haven, CT

PathMaster is a prototype cell-image database developed to serve as a pilot Next Generation Internet (NGI) testbed. PathMaster extracts mathematical features from cytology images and uses those features to index and retrieve cell images from an image database. First, the images are segmented, to facilitate differentiation of the cellular components: cytoplasm, nucleus and nucleolus. A semi-automated tool was developed to aid in the image segmentation process. This tool utilizes optical density thresholds to help distinguish the various cellular components. A total of 2390 features are calculated for each cell image including geometric, statistical, and nuclear texture features. A subset of these features is used to retrieve images from the image database.

Using the PathMaster application, a "simulated Internet" testbed was developed, using a cluster of parallel workstations, to explore a variety of Quality of Service strategies. Five strategies were studied: first in first out queuing, priority queuing, weighted fair queuing, weighted random early detection and weighted fair queuing with traffic shaping. All the strategies perform better than FIFO. Traffic shaping may help protect critical applications from bandwidth hogs.

A cytology image database was developed containing 1,477 thyroid aspirate cell images and 2,990 lymph node touch prep cell images. A variety of nearest neighbor techniques were evaluated for indexing the database but no technique was clearly superior. Images from the correct diagnostic classes were obtained within the top five hits 87% (thyroid) and 70% (lymphoid) of the time. Pilot testing of linear discriminant, logistic regression and random forest techniques for image classification were also explored.